

# NeurIPS 2020 Notes

## Vancouver, BC, Canada

Sergey Kastryulin\*  
[snk4tr@gmail.com](mailto:snk4tr@gmail.com)

December 2020

## Contents

<b>1 Tuesday, December 8th</b>	<b>4</b>
1.1 Session: Poster Session 1 . . . . .	4
1.1.1 MRI Banding Removal via Adversarial Training [6] . . . . .	4
<b>2 Wednesday, December 9th</b>	<b>6</b>
2.1 Session: Deep Learning . . . . .	6
2.1.1 Spotlight: What if Neural Networks had SVDs? [17] . . . . .	6
2.1.2 Spotlight: Triple descent and the two kinds of overfitting: where why do they appear? [5] . . . . .	7
2.2 Session: Social/Adversarial Learning . . . . .	8
2.2.1 Spotlight: Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses [22] . . . . .	8
2.3 Poster Session 4 . . . . .	9
2.3.1 Experimental design for MRI by greedy policy search [1] . . . . .	9
<b>3 Thursday, December 10th</b>	<b>12</b>
3.1 Poster Session 6 . . . . .	12
3.1.1 Differentiable Augmentation for Data-Efficient GAN Training [26] . . . . .	12
3.1.2 Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning [9] . . . . .	13
3.1.3 Noise2Same: Optimizing A Self-Supervised Bound for Image Denoising [25] . . . . .	14
3.1.4 3D Self-Supervised Methods for Medical Imaging [23] . . . . .	15
<b>4 Friday, December 11th</b>	<b>16</b>
4.1 Workshop: Workshop on Deep Learning and Inverse Problems . . . . .	16
4.1.1 Uncertainty Quantification in Deep MRI Reconstruction [7] . . . . .	16
4.1.2 Model Adaptation for Inverse Problems in Imaging [8] . . . . .	17
4.1.3 fastMRI . . . . .	18

---

\*<https://github.com/snk4tr>

<b>5 Saturday, December 12th</b>	<b>23</b>
5.1 Workshop: Medical Imaging Meets NeurIPS . . . . .	23
5.1.1 DeepSim: Semantic similarity metrics for learned image registration [4] . . . . .	23
5.1.2 Using StyleGANs for Visual Interpretability of Deep Learning Models on Medical Images paper . . . . .	24
5.1.3 FastMRI keynote: Fast(er) MRI: a radiologist’s perspective . . . . .	25
<b>6 Acknowledgements</b>	<b>26</b>

This document contains notes I took during the events I managed to make it to at NeurIPS 2020. Please feel free to distribute it and shoot me an email at [snk4tr@gmail.com](mailto:snk4tr@gmail.com) if you find any typos or other items that need correcting.

Some papers have extensive descriptions and some not. It tells nothing about the level of importance of the ideas nor on the level of a paper. It is only a matter of my personal preference, which also is influenced by many other factors. The review is also very biased towards inverse problems in imaging and in medical imaging in particular because this is what I do.

# 1 Tuesday, December 8th

## 1.1 Session: Poster Session 1

### 1.1.1 MRI Banding Removal via Adversarial Training [6]

Talk by *Aaron Defazio*.

**Motivation:** even current SOTA reconstructions of cascades of U-Nets may produce horizontally aligned banding artefacts (fig. 1). Presence of such artefacts can be a major roadblock towards deploying MR reconstruction systems in production.



Figure 1: Examples of banding artefacts presented on a reconstruction using "cascade of U-Nets" type of approach. Some of the artefacts are pointed with errors but actually they are everywhere on the image.

**Note:** these artefacts are particularly evident for scans taken with 1.5T machines due to their additional noise.

**Observation:** the direction striking artefacts is aligned with the direction of Cartesian undersampling. Hence if undersampling direction is changed by 90 degrees, horizontal artefacts will turn into vertical ones.

**Method:** the full pipeline is presented on fig. 2. The main idea is to train a model with an adversarial component, where the discriminator network will distinguish between different directions of artefacts. For that, k-space can be transposed during the data preparation process (while the orientation of the image is preserved).

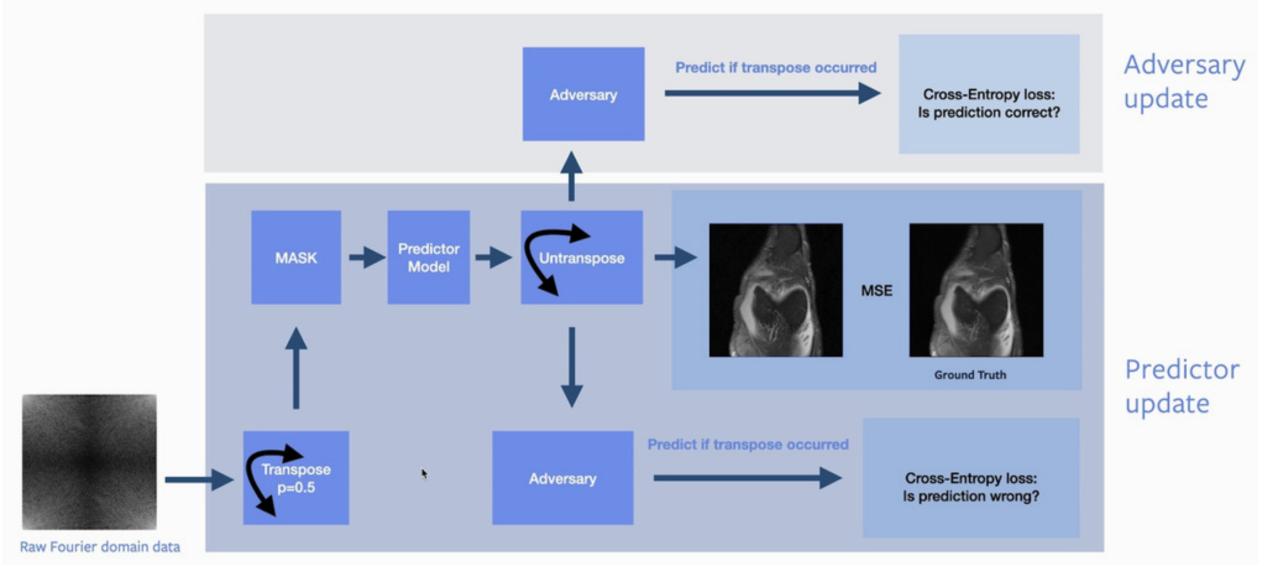


Figure 2: Illustration of the full pipeline. Initial raw k-space data can be transposed with  $p = 0.5$  and then fed into any reconstruction model. Next, transposed images are reverted back to the normal orientation and together with not transposed ones are fed into a discriminator network. Together with the initial supervised loss, adversarial loss is used to optimize reconstruction model. The proposed system can be trained end-to-end.

**Results:** the proposed technique produces outputs with little or no banding artefacts (fig. 3).



Figure 3: Illustration of results from the proposed artefact removal system. From left to right: fully sampled ground truth image (no artefacts), output of the standard model (artefacts are present), output of the proposed system trained with the adversarial component (less artefacts).

## 2 Wednesday, December 9th

### 2.1 Session: Deep Learning

This section appeared to be quite diverse on topics and it was hard to understand the flow. However, there still were a couple of talks that I found to be interesting.

.....

#### 2.1.1 Spotlight: What if Neural Networks had SVDs? [17]

Talk by *Alexander Mathiasen*.

**Motivation:** the SVD allows faster computation of some matrix operations commonly used by Neural Networks:

- Spectral Normalization [18] - a very popular technique to stabilize training of Generative Adversarial Networks (GANs)
- matrix inverse and Cayley transform - essential to train flow-based models (type of generative models, alternatives to GANs) [12]
- other applications such as matrix determinant and matrix exponential are less practical

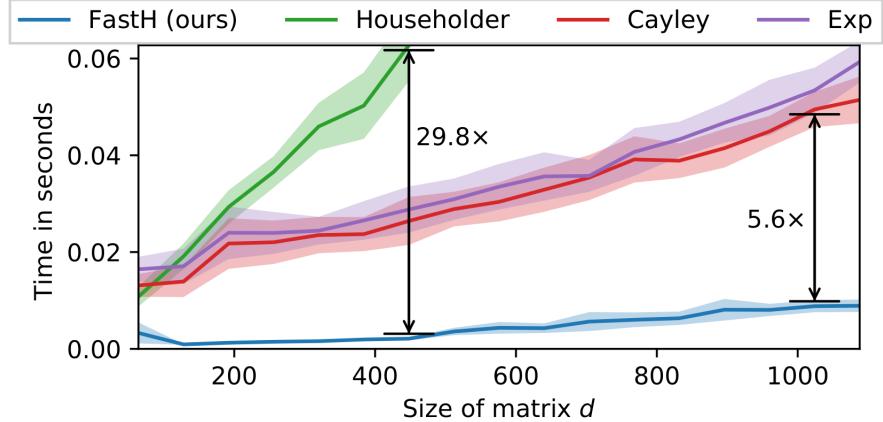
**Observation:** computation of SVD can be included in the computation of gradient descent. In this case, getting SVD results from a layer will take no time (because values are already computed when a user asks for them).

**Problem:** SVD cannot be directly formulated as a part of gradient descent step:

$$\hat{M} \neq M - \eta \nabla_M \quad (1)$$

**Method:** develop the Fast Householder Multiplication (FastH) method, which is a modification of the Householder transformation that supports parallel execution to better utilize GPU resources and includes some linear algebra tricks.

**Conclusions:** new method is fast on small matrices and also his complexity grows very slowly with grows of size of the decomposed matrix:

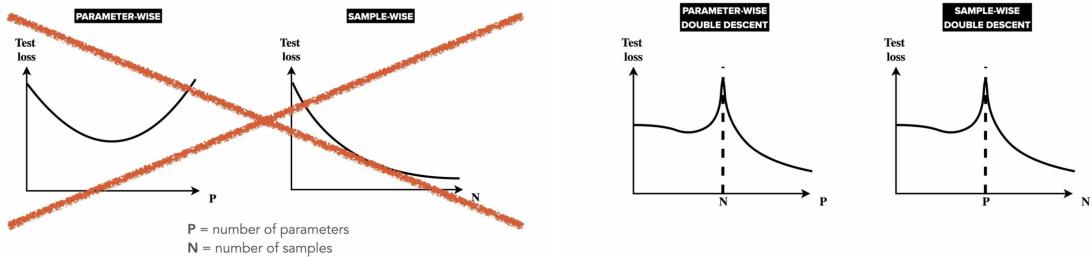


**Key takeaway:** code is released on [GitHub](#), try it out.

### 2.1.2 Spotlight: Triple descent and the two kinds of overfitting: where why do they appear? [5]

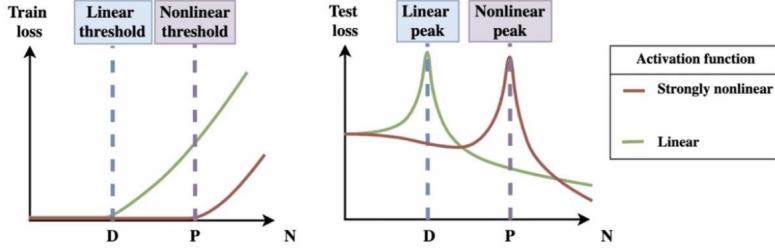
Talk by *Stéphane d'Ascoli*.

**Motivation:** when we increase the number of parameters in a machine learning model we expect loss function to decrease and then increase following the U-shape curve following the systematic bias-variance trade-off. When we increase the dataset size we expect train loss value to diminish and generalization to improve. However, this not something that happens in real Deep Learning.



In this work authors focus on the sample-wise double descent curve. It was discovered in 2018 by Belkin et. al. [3] and the over-fitting peak happen to co-inside with the interpolation threshold whether training loss becomes non-zero. Much earlier similar phenomena was discovered for linear models (perception) [LeCun et. al., 1991].

**Observation:** the non-linear peak occurs when number of training parameters  $P$  is of the same order as the number of training examples  $N$ . However, in linear model the over-fitting peak occurs when  $N$  is of the same order as input dimension  $d$ . Morevoer, in case of noisy regression task with weakly non-linear activation functions both peaks can be observed simultaneously.



**Goal of the work:** understand what cause these peaks.

**Method:** vary degree on non-linearity of activation functions, analyse eigenvalues of Gramm matrices of tested models (fully connected).

**Conclusions:**

- Linear peak is caused by the variance due to the noise in the labels
- Non-linear peak occurs due to initialization of random features

The consequence of this is that even if there is no noise in labels, non-linear peak when  $N$  equal to  $P$  still occurs, whereas the linear peak vanishes.

**Key takeaway:** this explains why in Deep Learning we experience over-fitting even there is no noise in the dataset.

## 2.2 Session: Social/Adversarial Learning

### 2.2.1 Spotlight: Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses [22]

Talk by *Gaurang Sriramanan*.

**Motivation:** current SOTA PGD attack may not be effective due to the non-convex nature of many Machine learning tasks leading to correct response from the model under attack. Existing modifications such as performing PGD attack sequentially with respect to several targets introduces additional computational cost.

**Observation:** better selection of initial gradient direction leads to a stronger attack.

**Method:** introduce the new attack called Guided Adversarial Margin Attack (GAMA):

$$L = -f_\theta^y(\tilde{x}) + \max_{j \neq y} f_\theta^j(\tilde{x}) + \lambda \cdot \|f_\theta(\tilde{x}) - f_\theta(x)\|_2^2 \quad (2)$$

First two terms of equation 2 correspond to the maximum margin loss in probability space. Next, there is a smoothing term, which is represented as L2 distance between probability vectors of clean image and the perturbed image.

## Relaxation term to improve optimization

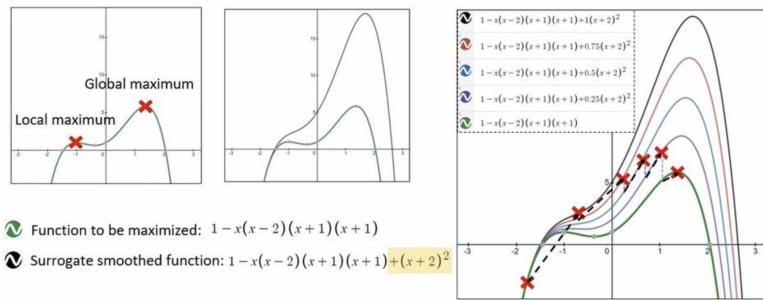


Figure 4: 1D example problem, which becomes more convex with introduction of the relaxation term. Loss surface becomes smoother hence leading to easier optimization and stronger attack.

**Observation:** introduction of a relaxation term (regularization) leads to smoother loss function (more convex problem) and hence better outcome of the attack.

**Side note:** the regularization term is added with  $\lambda$  parameter, which decays over optimization process and zeros-out at the end hence not effecting the ability of the system to reach an optimal solution.

**Results:** stronger attacks with lower number of iterations (100 iterations are considered for almost all experiments in the paper). GAMA attack is tested not only against plain models, but also against current SOTA defences.

**Key takeaways:** simple yet effective idea. The only disadvantage - only applicable for classification task, which limits its adoption in broader range of applications.

## 2.3 Poster Session 4

### 2.3.1 Experimental design for MRI by greedy policy search [1]

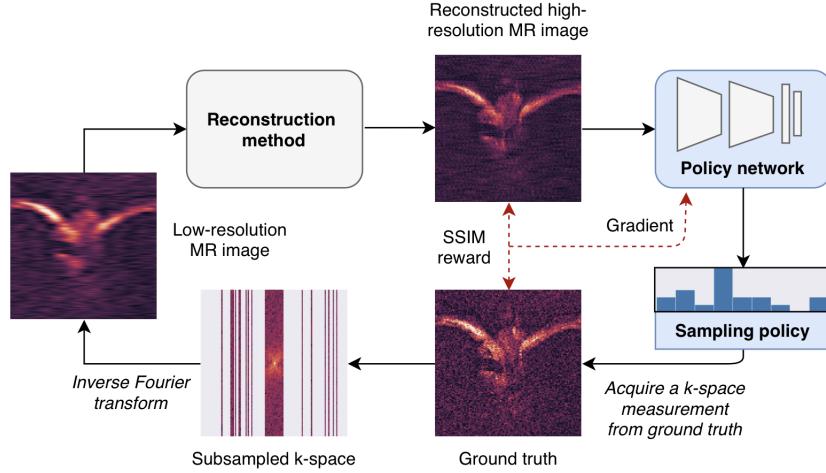
Presented by *Tim Bakker*

**Motivation:** move away from handcrafted sampling masks towards learned acquisition strategies.

**Observation:** the problem can be solved by joint optimization of some sampling scheme generator and the following reconstruction model. However, it leads to non-adaptive sampler: decision on which k-space lines to sample is made ones and not adjusted based on quality of recons made on partially sampled data. RL-based approach allows to make sampling scheme generation to be more adaptive.

**Method:** an initial subsampling of k-space is obtained from the ground truth image. The subsampled frequency domain is fed into a reconstruction method, which for neural network based

reconstructions starts with an inverse Fourier transform. This intermediate step results in a zero-filled reconstruction. The high-resolution reconstructed MR image is input to the policy network, which outputs a discrete probability distribution that represents the suggested sampling policy. A action is sampled from this policy, corresponding to a measurement of k-space. This measurement is simulated from the ground truth MR image, and the procedure is repeated until the acquisition budget is exhausted. The reward of an acquisition step is given by the improvement in Structural Similarity Index of the ground truth and reconstruction resulting from that acquisition.



**Side note:** one of the biggest practical disadvantages of the approach is that reconstruction model is applied after each action  $a$  meaning that several dozens of reconstruction may need to be performed before the desired resulting reconstruction is obtained.

**Results:** results were only estimated base on SSIM values, no manual comparison with application specialists is done.

- Based on SSIM values, Greed model performs better compared to other models. Learned policy was also compared with Random strategy that shares the initial mask with the other methods, and subsequently obtains a uniformly random measurement every acquisition step, similar to a simple VDS heuristic [16]
- The Greedy model is furthermore much less computationally expensive to train than any of the non-greedy models
- The Greedy model shows that it is adapting its predictions to individual MR images by outperforming the NA Oracle (selects as a measurement candidate in each acquisition step the column that leads to the greatest average SSIM improvement over the test dataset). This suggests that adaptivity is indeed a useful property for subsampling models to possess, and that this information can be learned in practice
- Method is verified for the single-coil case. It may be interesting to extend the work to more clinically relevant multi-coil case

**Q&A session with *Tim Bakker***

**Q:** How much is your learned policy better than some predefined one that people typically use?

**A:** We have not made this comparison yet but there is a feeling that only application specialist will see the difference (remark: they do not have application specialists/clinical scientists in their research group)

**Q:** How much reconstructions from different policies are visually different?

**A:** There is a huge gap in SSIM values but it is harder estimate visual difference (remark: seems like they have not done a proper visual evaluation)

**Q:** In the paper you tried to learn a policy with a fixed recon model. Have you consider a joint learning?

**A:** We have not. However, we understand that joint optimization may lead to better results. It should be confirmed though (remark: our current SDF learning approach is joint and hence may be advantageous).

**Q:** Have you tried more clinically relevant (lower) acceleration factors?

**A:** Not really. We stick to the fastMRI accelerations of x4 and x8 and then proceeded with higher accelerations (x32) because we were interested what will happen with the agent and reconstructions.

**Q:** Have you considered non-cartesian sampling schemes?

**A:** Very little, it would be honest to answer no here.

**Q:** What are your plans? Are you going to develop the method further? (remark: the question was asked with an intent to find out about active sampling. However, I did not mention it explicitly)

**A:** Not really. Currently, we are exploring options for clinical application and it draws all our attention.

### 3 Thursday, December 10th

#### 3.1 Poster Session 6

This poster session contained quite some papers related to the medicine. Self-supervision seem to be the general trend these days.

---

##### 3.1.1 Differentiable Augmentation for Data-Efficient GAN Training [26]

Presented by *Shengyu Zhao*.

**Motivation:** collecting datasets is expensive. However, GANs trained on limited datasets suffer from low performance because it becomes easier for discriminator  $D$  to just memorize training set (overfitting). Figure 5 illustrates this phenomena.

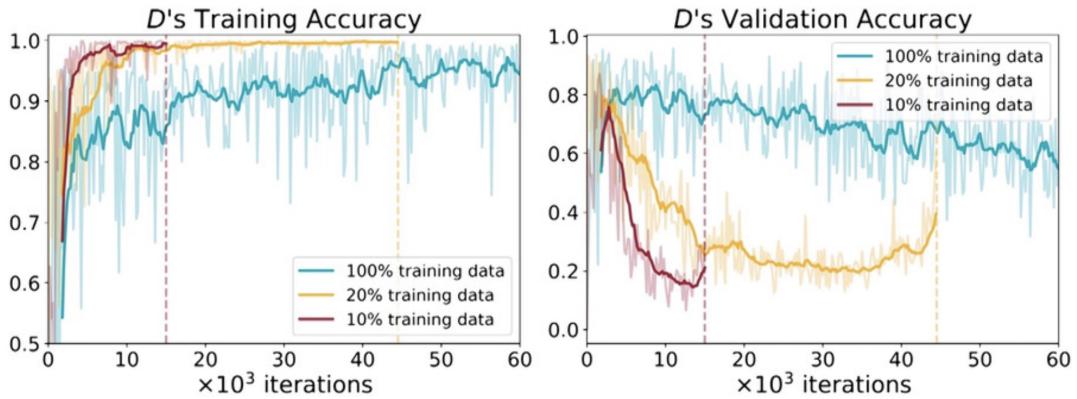


Figure 5: Accuracy of the discriminator  $D$  model rapidly increases when number of training samples becomes smaller. In the 20% case training accuracy rapidly saturates while validation accuracy keeps decreasing until collapse. This phenomena becomes more severe with only 10% of data available.

**Observation:** data augmentation can increase variability of data in case of limited datasets.

#### Problems:

- if augmentations are only introduced to the real images, outputs generated by the generator  $G$  will suffer from artefacts similar to the used augmentations
- if augmentations are introduced to both real and fake images but used only to train  $D$ , it leads to even worse results because  $G$  and  $D$  now are trained for different tasks

**Proposal:** introduce differentiable augmentations to modify real and fake images for both  $G$  and  $D$ . Consider the following basic GAN definition (equation 3):

$$\begin{aligned} L_D &= \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[f_D(-D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}[f_D(D(G(\mathbf{z})))] \\ L_G &= \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}[f_G(-D(G(\mathbf{z})))] \end{aligned} \quad (3)$$

Equation 3 is changed to introduce differentiable augmentation  $T$  as a part of the optimization problem (equation 4):

$$\begin{aligned} L_D &= \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[f_D(-D(\textcolor{red}{T}(\mathbf{x})))] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}[f_D(D(\textcolor{red}{T}(G(\mathbf{z}))))] \\ L_G &= \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}[f_G(-D(\textcolor{red}{T}(G(\mathbf{z}))))] \end{aligned} \quad (4)$$

**Results:** Even trained on 20% or 10% of data using this method results are reasonable, FID [10] scores are lower (see figure 6).

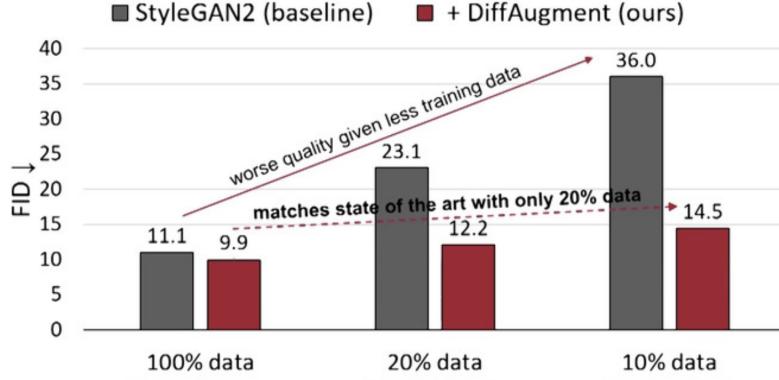


Figure 6: Illustration of advantages of using differentiable augmentation on when size of available dataset is limited. Gray bins represent FID scores for GAN trained without augmentations, red bins represent the same baseline model but trained using differential augmentations.

**Key takeaways:** differentiable augmentations is in general a great idea, which is extremely useful in practice (proven by lots of independent experiments) for training GANs and other frameworks. Also code is released [on GitHub](#).

### 3.1.2 Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning [9]

Presented by *Jean-Bastien Grill* (also oral at the Track 27 - Unsupervised/Probabilistic)

**Motivation:** improve self-supervised pretraining of CNNs to provide better initialization for further fine-tuning on the end task.

**Observation:** model can be trained not using labels from the training set. Instead, they can be generated by the network itself (its replica), which will allow to train without negative labels (potentially better performance).

## BYOL

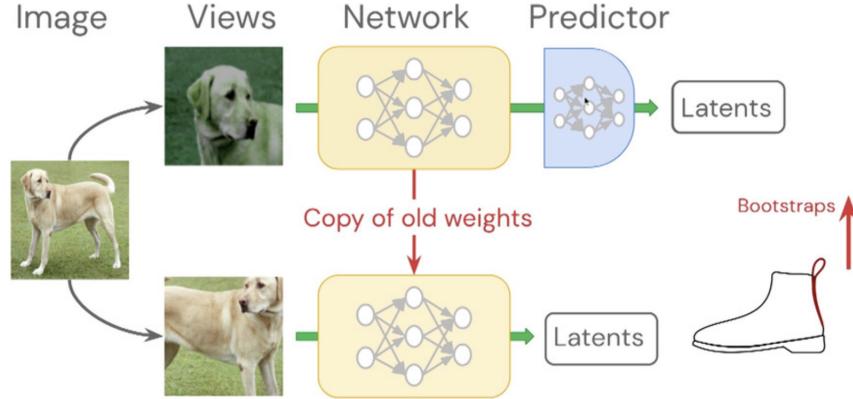


Figure 7: Main elements of the BYOL method. First, two different views are extracted from an image from the training set. Second, on the views is fed into the feature extractor and next to the predictor. The predictor outputs latents, which should match bootstrapped outputs of feature extractor replica.

**Method:** the main features of the method are illustrated on figure 7.

There are several key ingredient of this method:

- Carefully tuned image transformations
- Presence of the target network (inspired by Reinforcement Learning)
- Additional predictor on online network (makes training more stable)

**Results:** current SOTA, more resilient to transformation and batch size than contrastive methods.

**Key takeaways:** need to try self-supervision in other tasks (may be start with a simpler method). Code is [released on GitHub](#).

### 3.1.3 Noise2Same: Optimizing A Self-Supervised Bound for Image Denoising [25]

Presented by *Yaochen Xie*.

**Motivation:** enhance previous SOTA self-supervised denoising model Noise2Self [2] by solving its main problem - utilization of information about center pixel without knowing the noise model.

**Related work:** Noise2Noise [15] compared to the proposed approach requires more supervision with noisy-noisy pairs. Here, only unpaired noisy images are available. In Noise2Void [13] and Noise2Self [2] **blind spot denosing** is proposed. Here, some masks are applied and then each

pixel is denoised with this context. The biggest disadvantage of this approach is that only surrounding pixel are utilized while the information from the pixel being denoised is not used. This leads to lower performance. Some methods to address it were proposed by they rely on noise model, which is unknown in most cases.

**Observation:** one term in the loss function of the Noise2Self model can be bound to be zero instead of assuming it.

**Method:** in general, method follows the Noise2Self framework. The only difference is the objecting function. It is changed to meet zero-value requirement instead of assuming it.

$$\mathbb{E}_x \|f(\mathbf{x}) - \mathbf{x}\|^2 = \mathbb{E}_{x,y} \|f(\mathbf{x}) - \mathbf{y}\|^2 + \mathbb{E}_{x,y} \|\mathbf{x} - \mathbf{y}\|^2 - 2 \langle f(\mathbf{x}) - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle$$

**Part of the**  
Self-supervised loss  
in Noise2Same

*Supervised loss*

We **bound this term** in  
a self-supervised  
fashion, instead of  
assuming “J-invariant”  
and making it zero

**Results and takeaways:** SOTA results compared to other self-supervised/weakly supervised methods, code is [released on GitHub](#).

### 3.1.4 3D Self-Supervised Methods for Medical Imaging [23]

Presented by *Aiham Taleb*.

**Motivation:** most of known self-supervised method are developed for 2D data. However, there is a way to utilize 3D spacial information, which is often available in medical imaging.

**Method:** adapt 2D self-supervised pretraining methods to 3D. The following methods are adapted:

- Contrastive Predictive Coding
- Rotation Prediction
- Relative Patch Location
- Jigsaw Puzzle Solving
- Exemplar Networks

**Results and takeaways:** not much novelty but code for the proposed 3D methods is [released on GitHub](#).

## 4 Friday, December 11th

### 4.1 Workshop: Workshop on Deep Learning and Inverse Problems

Even though this workshop is dedicated to inverse problems in broad sense of this term, the majority of presentations are somehow related with imaging. Moreover, significant fraction of them is either dedicated to MRI reconstruction or at least somehow mention it. This observation underlines consensus of the research community on the importance of this problem.

#### 4.1.1 Uncertainty Quantification in Deep MRI Reconstruction [7]

Presented by *Vineet Edupuganti*.

**Motivation:** accelerated MRI reconstruction is unreliable, especially when generative models are applied. Need to estimate uncertainty of a reconstruction model to help radiologist to avoid mistakes.

**Background:** most of existing approaches (Deep Ensembles [14], VAE sampling [20]) for epistemic uncertainty estimation leverage one or another form of the Monte Carlo sampling: several samples are utilized to compute statistics, which are then represented as results. One example of this approach is to sample from VAE and compute variance of results. However, it is impossible to compute bias without ground truth image using this method.

**Method:** authors propose to utilize Stein's Unbiased Risk Estimator (SURE) [24] to compute bias-free uncertainty estimation without reference image. For that, the following derivation is performed.

Given the ground truth image  $x_0$ , the zero-filled image can be written as

$$x_{zf} = x_0 + v \quad (5)$$

where  $v$  is noise. Now considering reconstruction  $\hat{x}$  with dimension  $n$ , one can expand test MSE as

$$\mathbb{E}\|\hat{x} - x_0\|^2 = \mathbb{E}\|x_0 - x_{zf} + x_{zf} - \hat{x}\|^2 = -n\sigma^2 + \mathbb{E}\|x_{zf} - \hat{x}\|^2 + 2Cov(x_{zf}, \hat{x}) \quad (6)$$

Since  $x_0$  is not present in 6, we see that SURE serves as a surrogate for MSE even when the ground truth is unknown. A key assumption behind SURE is that the noise process  $v$  that relates the zero-filled image to the ground truth is normal, namely  $v \sim N(0, \sigma^2 I)$ . With this assumption SURE is formulated as follows

$$SURE = -n\sigma^2\|\hat{x} - x_{zf}\|^2 + \sigma^2 tr(\partial\hat{x}/\partial x_{zf}) \quad (7)$$

**Important note:** previous assumption may not hold in practice.

Authors show that when error in the output reconstruction is not large, and as a result  $\sigma^2$  can be estimated as

$$\sigma^2 = \|\hat{x} - x_{zf}\|^2/n \quad (8)$$

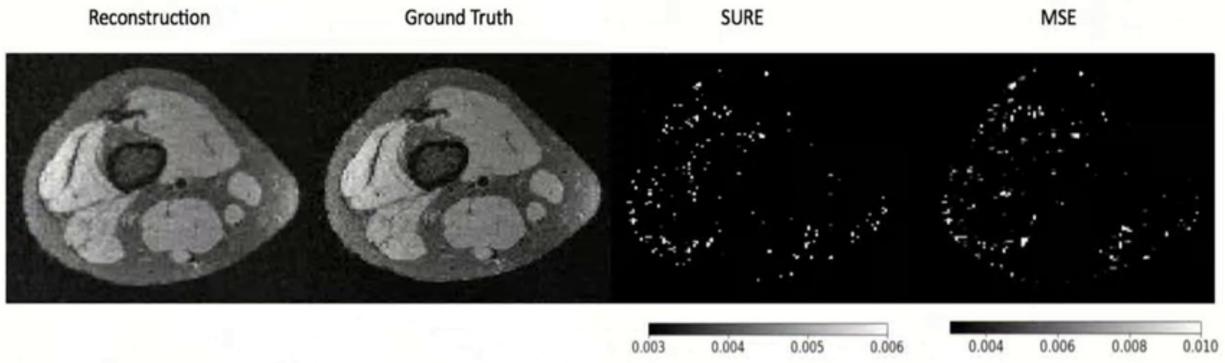
This assumption allows to formulate SURE as follows

$$SURE = \sigma^2 \text{tr}(\partial \hat{x} / \partial x_z f) \quad (9)$$

**Important note:** this assumption only holds for low acceleration rates and artefact-free reconstruction models. Since this method is developed for evaluation of potentially artefact-prone models, this assumption looks quite strong.

Further investigations are targeted on acceleration of computations of trace of the end-to-end network Jacobian, which otherwise can be problematic. Authors also approach the problem of the requirement for MRI noise to follow standard normal distribution.

**Results and takeaways:** the main contribution of this work is to apply SURE to the uncertainty estimation task in the domain of medical imaging, which has not been done before. An example of high correlation between estimated SURE values and computed MSE error is illustrated on figure ??.



Pixel wise SURE and MSE maps for a reference slice

#### 4.1.2 Model Adaptation for Inverse Problems in Imaging [8]

Presented by *Rebecca Willett*.

**Motivation:** having measurements  $y$  and a neural network  $f_\phi$  trained with some forward model  $A_0$ , make high quality reconstructions  $\hat{x}$  if forward model used during inference  $A_1$  is different from the one used for training ( $A_0$ ). The scheme of the task setup is represented on fig. 8. Authors call this phenomenon a *model drift*.

**Example:** train a fast MRI network with one acceleration rate but perform inference on some different acceleration rate. Even if the acceleration rate is lower, there is a domain shift that makes the network less capable to produce a high-quality estimate.

**Definitions:**

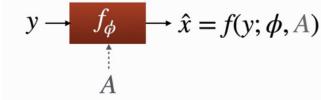


Figure 8: Neural network with parameters  $\phi$  (and optionally  $A$ ) inputs  $y$  and outputs an estimate  $\hat{x}$ .

- **Distribution drift:**  $p(X, Y)$  changes *in unknown way* between train and deployment
- **Model drift:**  $p(X, Y)$  changes *in known or partially known way* between train and deployment
- **Domain adaptation:** first train with many samples  $(x_i^{(0)}, y_i^{(0)}) \sim p_0$ , then adapt using few samples from  $(x_i^{(1)}, y_i^{(1)}) \sim p_1$
- **Model adaptation:** first train with many samples  $(x_i^{(0)}, y_i^{(0)}) \sim p_0$  and, then adapt using few samples from  $(?, y_i^{(1)}) \sim p_1$  (images are not known)

**Method:** train a reconstruction network for a known forward model then adapt to a new forward model without access to ground truth images, and without knowing the exact parameters of the new forward model.

Proposed approaches are:

**Parametrize and Perturb:** use calibration data to *perturb* parameters of the original reconstruction network. In practice it means conversion of  $f_0$  to some  $f_1$  that will be able to handle  $A_1$ .

**Condition and Correct:** train a new (correction) network  $h_\theta$  to map  $A_1$  to  $A_0$  and then apply  $f_\phi$ .

**Results:** authors show high-quality results (fig. 4.1.2, 11, 12, 13), which confirm that their methods allow to increase the quality of resulting images if model drift takes place. More concretely:

- Condition and Correct method, in general, works a little bit better.
- The most significant visual improvements are presented for the motion correction problem.

#### 4.1.3 fastMRI

Presented by *Larry Zitnick*.

**Motivation:** describe the fastmMRI problem, present a network by organizers of the fastMRI challenge, describe model evaluation process

**Note:** for some reason problem of the fastMRI challenge 2019 are discussed. Not sure why authors did not also include discussion of this year's problem.

**Problem:** reconstructions from the model have less noise hence look unnatural to radiologists.

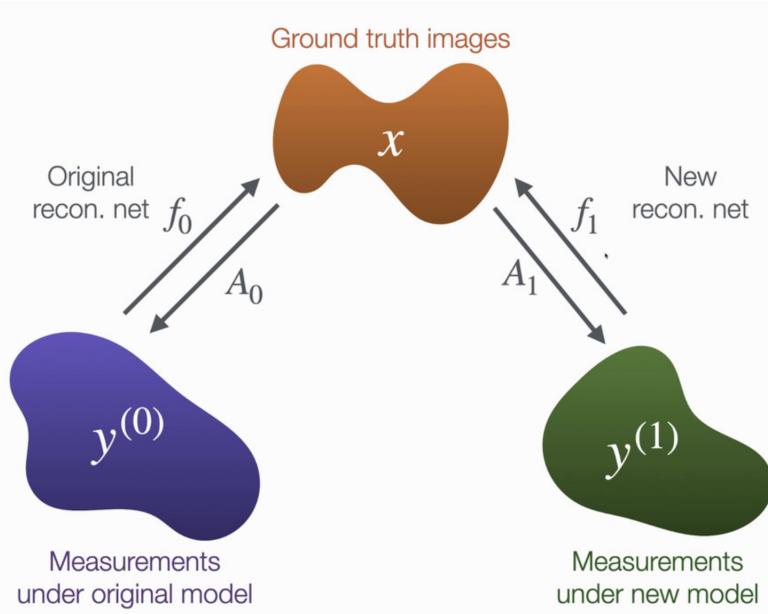


Figure 9: Illustration of the Parametrize and Perturb approach.  $f$  is perturbed to be able to handle  $A_1$  during the inference time.

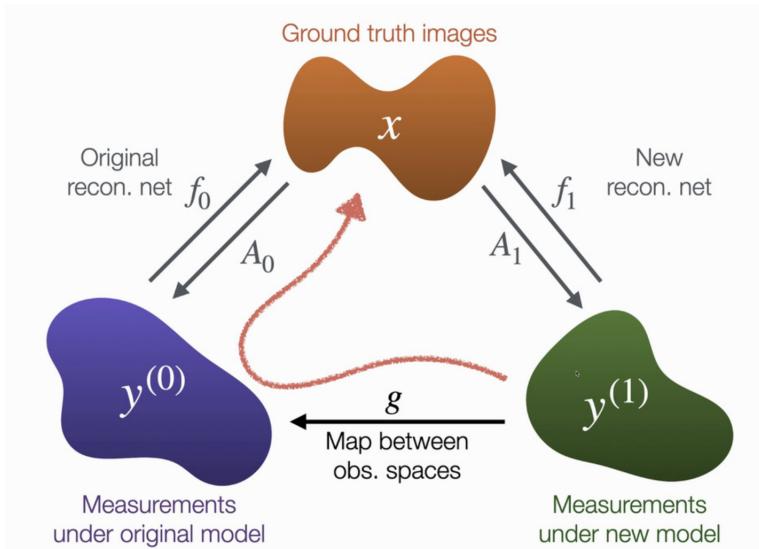


Figure 10: Illustration of the Condition and Correct approach. Output of the  $A_1$  is made similar to  $A_0$  and then  $f_\phi$  is applied

**Solution:** add noise to the model recons (fig. 14). This approach worked good for them (radiologists are satisfied).

## Performance with unknown $A_1$

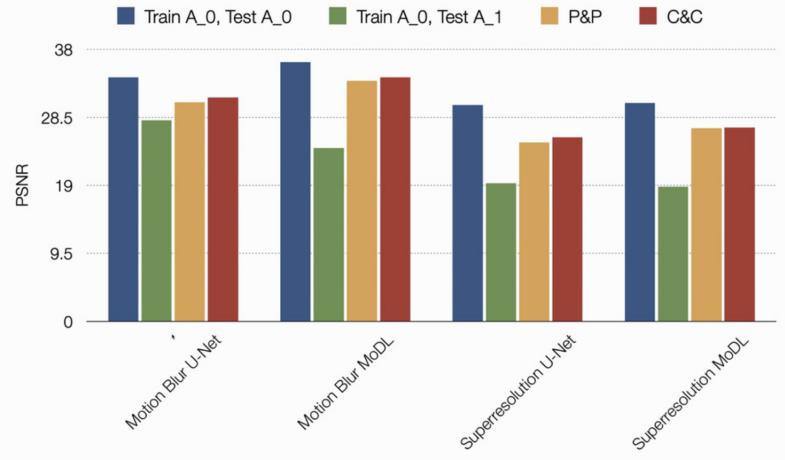


Figure 11: Comparison between performance of the reconstruction model in case if  $A_1$  is unknown (realistic setting) on various inverse problems.

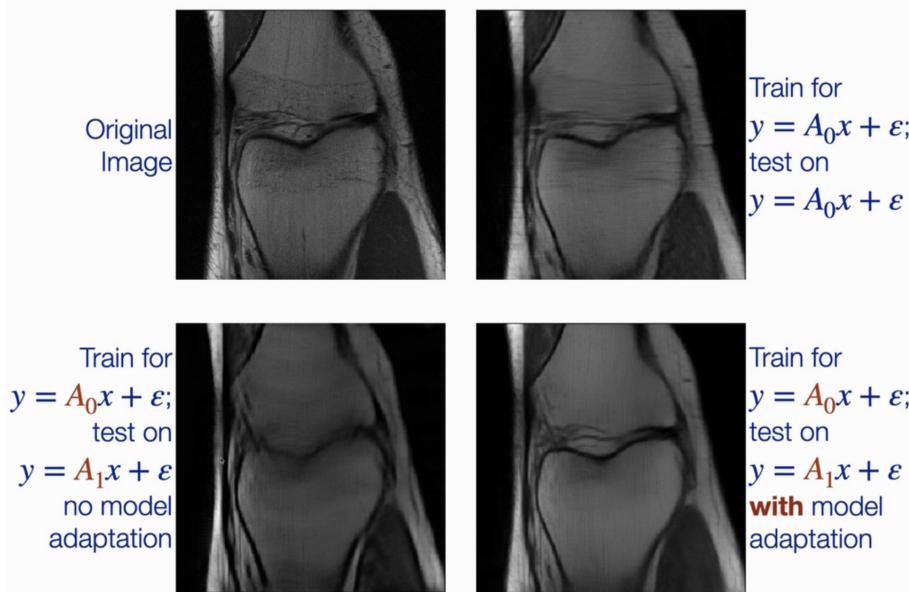


Figure 12: Example of the performance of the proposed method for the fast MRI reconstruction problem

**Observation:** authors experienced horizontal banding artefacts (fig. 15) in their multi-coil 4x reconstructions. Reason - insufficient sampling pattern. K-space is sampled vertically (Cartesian sampling), which results in higher resolution in vertical dimension compared to the horizontal dimension and hence causing horizontal artefacts (explanation for the presenter).

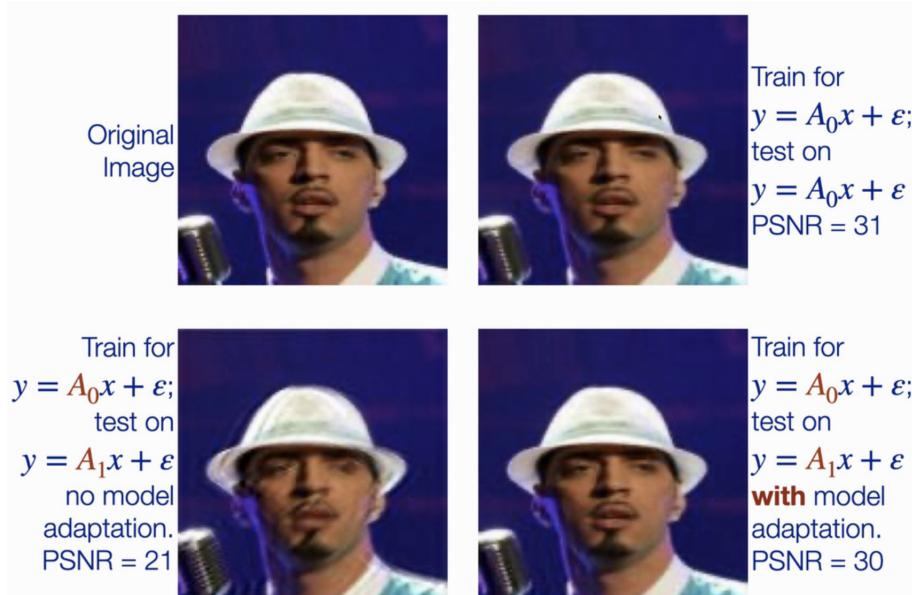


Figure 13: Example of the performance of the proposed method for the general domain motion correction problem

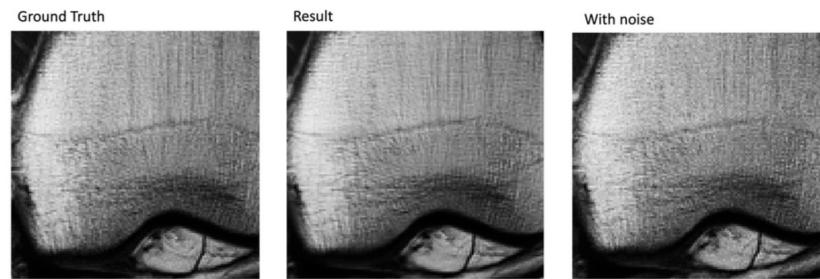


Figure 14: Example of high-quality but oversmoothed result from the reconstruction model.

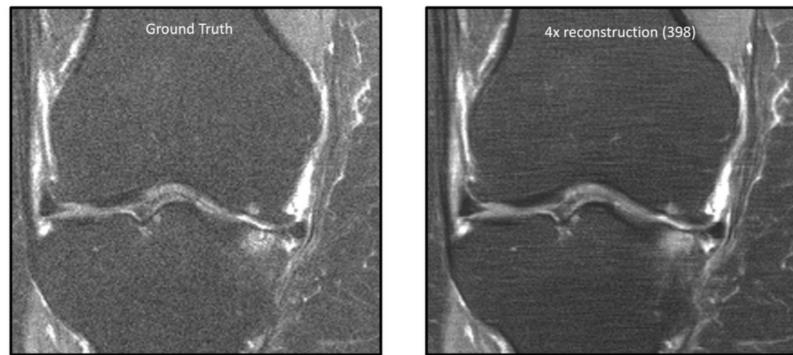


Figure 15: Horizontal binding artefacts caused by insufficient sampling pattern.

**Method:** most of the talk very similar to the same talk from the last year. The new part was the interchangeability study, which authors performed using 4x acceleration (3T) and 108 patients. Reconstructions were reviewed by 6 radiologists according to the following process:

- Radiologist reviews a scan
- Waits one month
- Same radiologist reviews the scan
- Separate radiologist compares the two reports

**Results:** in the previously described setting:

- New reconstructions are diagnostically interchangeable with traditional MRIs
- Radiologists could not distinguish which images were produced with AI and which came from the slower traditional scans
- All examiners rated the AI-generated images as being higher in quality than traditional scans

Note that all these results are obtained only for a single anatomy (knee) and for a limited number of protocols (only the ones that are present in the fastMRI knee dataset).

## 5 Saturday, December 12th

### 5.1 Workshop: Medical Imaging Meets NeurIPS

#### 5.1.1 DeepSim: Semantic similarity metrics for learned image registration [4]

Presented by *Steffen Czolbe*.

**Motivation:** having a decent solution of image registration as an ill-posed problem is heavily dependent on the presence of a meaningful distance metric  $D$  (fig. 16).

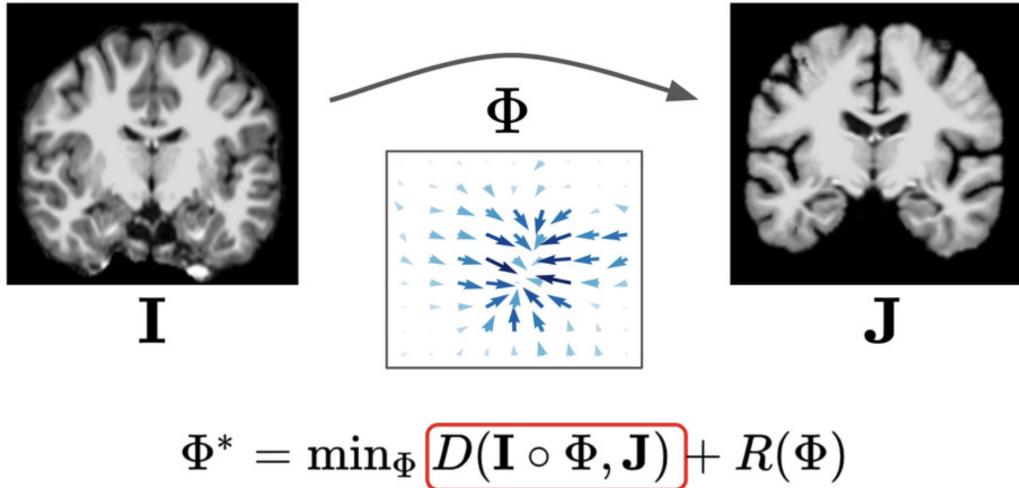


Figure 16: Medical image registration problem setting. Selection of a distance metric  $D$  plays a key role.

**Observation:** typically, the distance in image registration task is computed on patches as pixel-wise cosine distance (eq. 10) between reconstruction and some generated reference. The problem here is that pixel-wise esitimations are typically not very accurate and can be improved by replacing pixel with some more high-level features.

$$NCC_{patch}(A, B) = \frac{\langle A - \hat{A}, B - \hat{B} \rangle}{\|A - \hat{A}\| \|B - \hat{B}\|} \quad (10)$$

**Method:** pretrain some feature extractor on a proxy task withing the same imaging domain (segmentation), use its features as an enriched substitutions of raw pixel values (eq. 11).

$$DeepSim_{patch}(A, B) = \frac{\langle F(A), F(B) \rangle}{\|F(A)\| \|F(B)\|} \quad (11)$$

**Results:** as claimed:

- High registration accuracy across multiple datasets
- The metric is general and can be applied to different modalities and algorithms Translations results are smoother

### 5.1.2 Using StyleGANs for Visual Interpretability of Deep Learning Models on Medical Images [paper](#)

Presented by *Kathryn Schutte*.

**Motivation:** intractability of Machine Learning models is important, especially in medical imaging. The goal is to improve intractability by proposing an alternative go the currently popular GradCAM [21] approach.

**Observation:** currently used methods (for example GradCAM) only highlight regions of an image that has the highest contribution to the result that a model outputs. Instead, one could generate an image view that would lead to another output. Example is shown on the fig. 17.

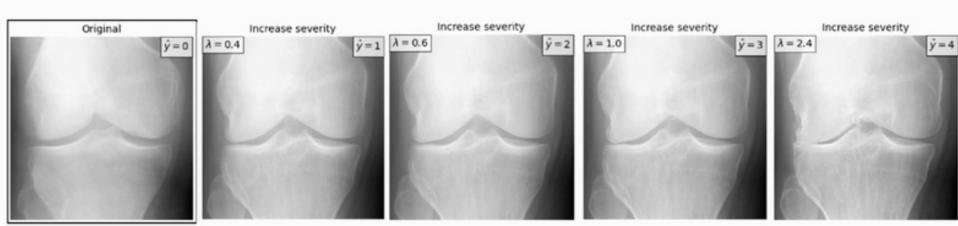


Figure 17: Illustration changes of image view that could lead to a different model output. In this case classification task is considered and the initial output of the model is 0 (no pathology). However, one can have a hard time understanding why the prediction is so. For that, the proposed method generates transformations of the initial image that would lead to the alternative classification results.

**Method:** the following pipeline is proposed:

**Step 1:** train a StyleGAN model [11]. StyleGAN is chosen because 1) it is able to generate high-quality images and 2) has an interpretable latent space vector in the bottleneck of the generator, which can be used to generate images with desired qualities (check out StyleGAN paper for more information).

**Step 2:** find the direction, which has the greatest impact on the model's output (in the particular case - on Osteoarthritis severity).

$$\tilde{f}(w) = \sigma(\alpha^T \mathbf{w} + \beta) \approx f(g(w)) \quad (12)$$

So if we approximate classifier fitting a logistic regression  $f$  to this latent space we can retrieve the direction of Osteoarthritis severity using  $\alpha$ . It means, that on this stage we first build a labeled dataset of latent vectors (otherwise they are hard to interpret).

**Step 3:** Encode the real images into the latent space (fig. 18).

**Results:** generated images seem to be reasonable, however the solution still looks very task-specific. Not clear whether the proposed method will give reliable results on other domains, especially considering the fact that GAN training is in general data-hungry and available datasets may not be enough. Also not clear how to transfer the approach for the regression task (if it is possible at all).

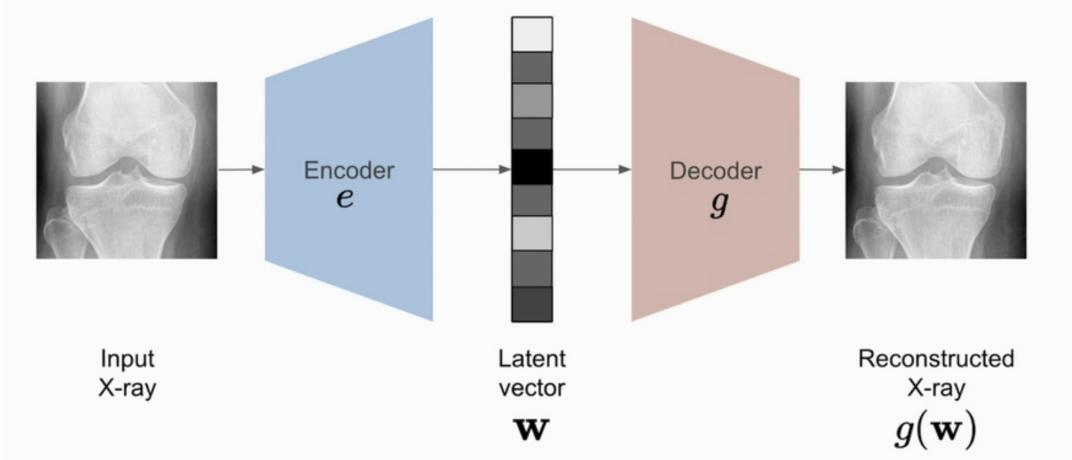


Figure 18: The last step of the proposed pipeline - encoding real images into latent vectors.

### 5.1.3 FastMRI keynote: Fast(er) MRI: a radiologist's perspective

Presented by *Yvonne Lui*.

**Motivation:** discuss results of this year's challenge, compare them with the ones from the previous year.

**Observation:** organizers claim that overall SSIM has a decent correlation with radiologist's rankings of challenge results. More concretely, when SSIM values are very different, their relative order well correlates with the order provided by manual evaluations with radiologists. However, when differences in values are small (3rd-4th digit), like we can commonly observe in top-3 of the leaderboard, relative order may be different. In particular, top-2 and top-3 results (according to their SSIM scores, difference in the 3rd digit) were flipped after evaluation from radiologists.

**Observation:** top-3 values of SSIM scores from 2020 challenge are much higher than corresponding values from the previous year. 2019 winner would not get into top-3 in 2020 with the same SSIM score. However, it is not clear whether such a direct comparison is fair due to the fact that 2019 and 2020 challenges utilized very different data (different anatomies, protocols).

**Observation:** last year's manual evaluation procedure was not aimed at finding the most accurate (from the anatomical perspective) reconstructions. Instead, radiologists only judged on the perceptual feeling of quality, which led to the problem illustrated on fig. 19. Here, none of the reconstructions was good enough to capture the important anatomical structure while radiologists ranked all these reconstructions to be high-quality.

To avoid the same problem this year, organizers revised the evaluation process and carefully selected several pathologies.

**New evaluation process:** radiologists were asked to rank reconstructions by assigning them scores from 1 to 4, where lower is better in for different criteria: presence of artefacts, sharpness

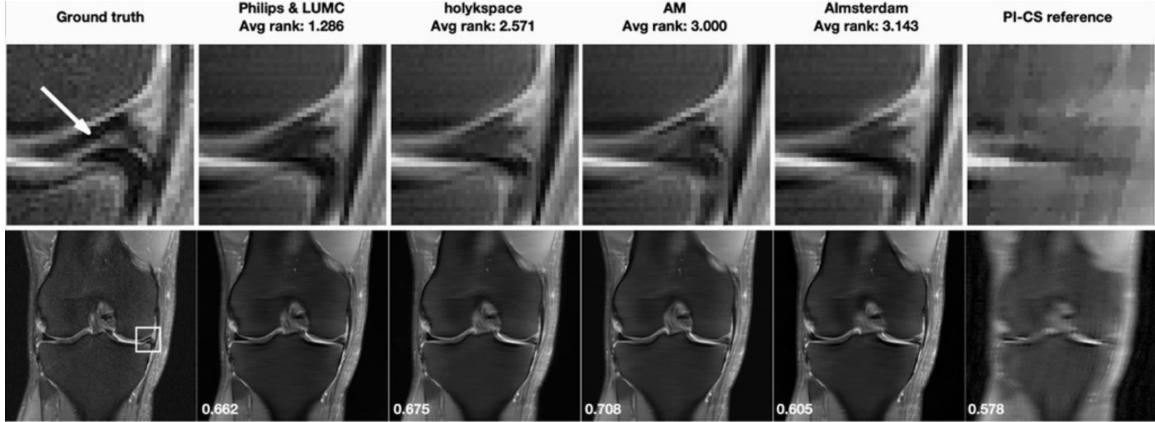


Figure 19: Example from the 2019 challenge, when all reconstructions failed to preserve an important anatomical structures but still were highly ranked by experts.

and contrast-to-noise ratio (CNR).

**Multi-coil 4x and 8x tracks:** this is very similar challenge to the 2019 version except from data and the evaluation process. The presentation of results showed that 8x acceleration is still too high for a potential clinical application with diagnostic purposes.

**Transfer track:** this track was aimed to estimate how good model trained on data from one vendor can be used on data from another vendors. In particular Siemens data was used for training and transferring was tested on GE and Philips data. Overall, Philips appeared to be closer to Siemens data resulting in (in general) better reconstructions. However, results greatly vary depending on contrast and particular anatomical structures.

**Note:** one of competitors this year also provided estimates of coil sensitivity maps (CSMs) as a part of their submission. This could help them to increase the overall quality (not clear whether it helped a lot though).

**Note:** presented mentioned the work from MRM (early 2020) on Gibbs deringing [19] where authors were able to create a system for Gibbs ringing removing. The artefacts were result of a partial Fourier transform. They did it by training a neural network on ImageNet dataset for the deringing task. Then this model was applied to MRI data and showed a decent performance.

## 6 Acknowledgements

I'd like to acknowledge the assistance of Mikhail Bortnikov who was helping me to not miss the most interesting parts of the conference, live poster presentations and tutorials.

## References

- [1] Tim Bakker, Herke van Hoof, and Max Welling. Experimental design for MRI by greedy policy search. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/daed210307f1dbc6f1dd9551408d999f-Abstract.html>.
- [2] Joshua Batson and Loïc Royer. Noise2self: Blind denoising by self-supervision. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 524–533. PMLR, 2019. URL <http://proceedings.mlr.press/v97/batson19a.html>.
- [3] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off, 2019.
- [4] Steffen Czolbe, Oswin Krause, and Aasa Feragen. Deepsim: Semantic similarity metrics for learned image registration. *CoRR*, abs/2011.05735, 2020. URL <https://arxiv.org/abs/2011.05735>.
- [5] Stéphane d’Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: Where why do they appear?, 2020.
- [6] Aaron Defazio, Tullie Murrell, and Michael P. Recht. MRI banding removal via adversarial training. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/567b8f5f423af15818a068235807edc0-Abstract.html>.
- [7] Vineet Edupuganti, Morteza Mardani, Shreyas Vasanaawala, and John Pauly. Uncertainty quantification in deep mri reconstruction, 2020.
- [8] Davis Gilton, Gregory Ongie, and Rebecca Willett. Model adaptation for inverse problems in imaging, 2020.
- [9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html>.
- [10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In

- Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. URL <http://papers.nips.cc/paper/7240-gans-trained-by-a-two-time-scale-update-rule-converge-to-a-local-nash-equilibrium.pdf>.
- [11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00453. URL [http://openaccess.thecvf.com/content\\_CVPR\\_2019/html/Karras\\_A\\_Style-Based\\_Generator\\_Architecture\\_for\\_Generative\\_Adversarial\\_Networks\\_CVPR\\_2019\\_paper.html](http://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html).
  - [12] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions, 2018.
  - [13] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void - learning denoising from single noisy images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2129–2137. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00223. URL [http://openaccess.thecvf.com/content\\_CVPR\\_2019/html/Krull\\_Noise2Void\\_-\\_Learning\\_Denoising\\_From\\_Single\\_Noisy\\_Images\\_CVPR\\_2019\\_paper.html](http://openaccess.thecvf.com/content_CVPR_2019/html/Krull_Noise2Void_-_Learning_Denoising_From_Single_Noisy_Images_CVPR_2019_paper.html).
  - [14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413, 2017. URL <http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf>.
  - [15] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2971–2980. PMLR, 2018. URL <http://proceedings.mlr.press/v80/lehtinen18a.html>.
  - [16] Michael Lustig, David Donoho, and John M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. doi: <https://doi.org/10.1002/mrm.21391>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.21391>.
  - [17] Alexander Mathiasen, Frederik Hvilsted, Jakob Rødsgaard Jørgensen, Anshul Nasery, and Davide Mottin. What if neural networks had svds?, 2020.
  - [18] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *6th International Conference on Learning Representations, ICLR 2018, New Orleans, LA, USA, April 30 – May 4, 2018*, pages 1–11. OpenReview.net, 2018. URL <https://openreview.net/forum?id=HJzDwHqQOg>.

- sentations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.* OpenReview.net, 2018. URL <https://openreview.net/forum?id=B1QRgziT->.
- [19] Matthew J. Muckley, Benjamin Ades-Aron, Antonios Papaioannou, Gregory Lemberskiy, Eddy Solomon, Yvonne W. Lui, Daniel K. Sodickson, Els Fieremans, Dmitry S. Novikov, and Florian Knoll. Training a neural network for gibbs and noise removal in diffusion mri. *Magnetic Resonance in Medicine*, 85(1):413–428, Jul 2020. ISSN 1522-2594. doi: 10.1002/mrm.28395. URL <http://dx.doi.org/10.1002/mrm.28395>.
  - [20] Xuming Ran, Mingkun Xu, Lingrui Mei, Qi Xu, and Quanying Liu. Detecting out-of-distribution samples via variational auto-encoder with reliable uncertainty estimation. *CoRR*, abs/2007.08128, 2020. URL <https://arxiv.org/abs/2007.08128>.
  - [21] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.74. URL <https://doi.org/10.1109/ICCV.2017.74>.
  - [22] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and Venkatesh Babu R. Guided adversarial attack for evaluating and enhancing adversarial defenses. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/ea3ed20b6b101a09085ef09c97da1597-Abstract.html>.
  - [23] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/d2dc6368837861b42020ee72b0896182-Abstract.html>.
  - [24] Ryan J. Tibshirani and Saharon Rosset. Excess optimism: How biased is the apparent error of an estimator tuned by sure? *Journal of the American Statistical Association*, 114(526):697–712, 2019. doi: 10.1080/01621459.2018.1429276. URL <https://doi.org/10.1080/01621459.2018.1429276>.
  - [25] Yaochen Xie, Zhengyang Wang, and Shuiwang Ji. Noise2same: Optimizing A self-supervised bound for image denoising. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/ea6b2efbdd4255a9f1b3bbc6399b58f4-Abstract.html>.
  - [26] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell,

Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/55479c55ebd1efd3ff125f1337100388-Abstract.html>.